PCT WORLD INTELLECTUAL PROPERTY ORGANIZATION
International Bureau

INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification 6 : **H04L 12/56**

A1

(11) International Publication Number: **WO 98/25381**

(43) International Publication Date: 11 June 1998 (11.06.98)

(21) International Application Number: PCT/US97/22620

(22) International Filing Date: 3 December 1997 (03.12.97)

(30) Priority Data:

60/032,124	5 December 1996 (05.12.96)	US
08/850,906	2 May 1997 (02.05.97)	US

(71) Applicant: 3COM CORPORATION [US/US]; 5400 Bayfront Plaza, Santa Clara, CA 95052 (US).

(72) Inventors: SHERER, William, Paul; 850 Pepperwood Drive, Danville, CA 94506 (US). CONNERY, Glenn; 655 S. Fair Oaks B301, Sunnyvale, CA 94086 (US).

(74) Agents: LeBLANC, Stephen, J. et al.; Townsend and Townsend and Crew LLP, 8th floor, Two Embarcadero Center, San Francisco, CA 94111 (US).

(81) Designated States: AU, CA, GB, JP, European patent (AT, BE, CH, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE).

Published

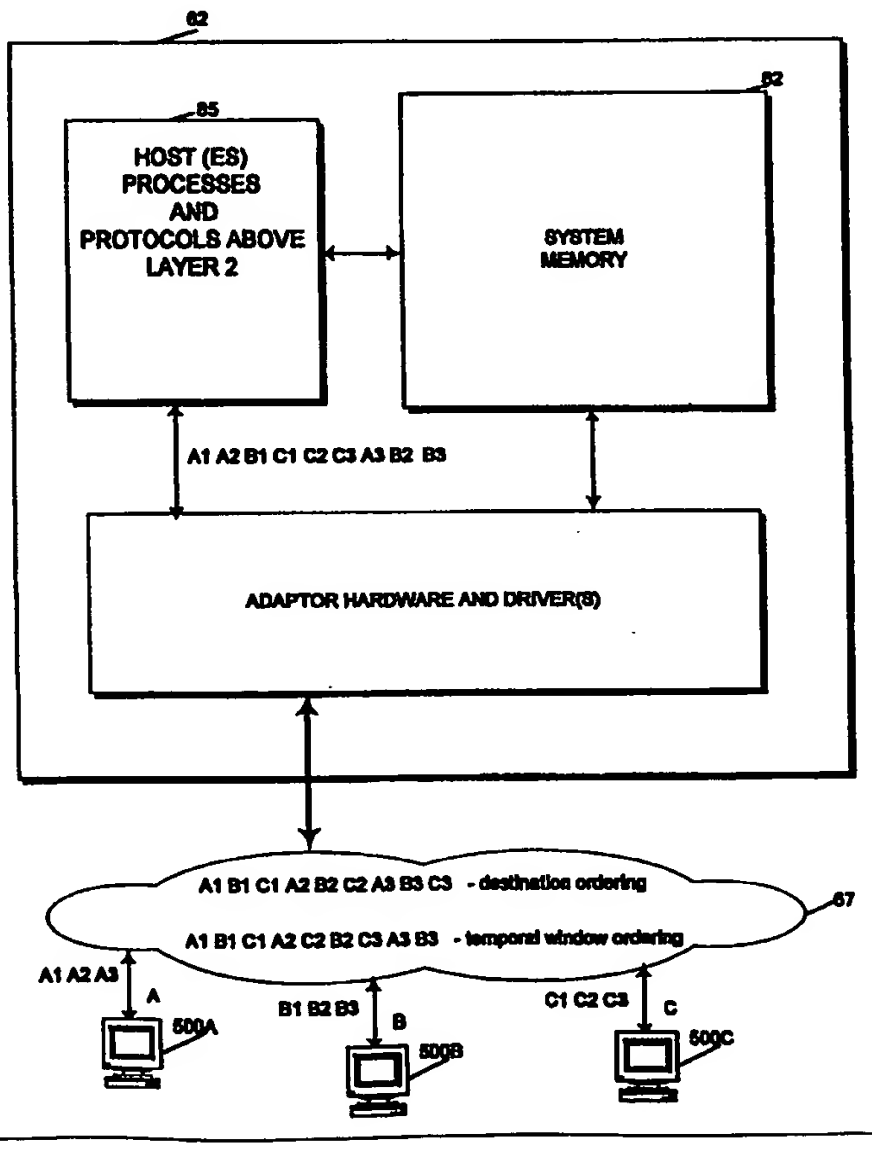
With international search report.

Before the expiration of the time limit for amending the claims and to be republished in the event of the receipt of amendments.

(54) Title: NETWORK ADAPTOR DRIVER WITH DESTINATION BASED ORDERING

(57) Abstract

A network transmitter (62) (driver or adaptor/driver combination) reorders packets received for transmission from a higher layer protocol (85) based on packet destinations. The invention reduces bottlenecks at the transmitter and potentially throughout network intermediate system. Destination based reordering may be accomplished through alternative methods, some alternatives taking into account the order at which packets to different destinations are queued by a higher layer protocol.



FOR THE PURPOSES OF INFORMATION ONLY

Codes used to identify States party to the PCT on the front pages of pamphlets publishing international applications under the PCT.

AL	Albania	ES	Spain	LS	Lesotho	SI	Slovenia
AM	Armenia	FI	Finland	LT	Lithuania	SK	Slovakia
AT	Austria	FR	France	LU	Luxembourg	SN	Senegal
AU	Australia	GA	Gabon	LV	Latvia	SZ	Swaziland
AZ	Azerbaijan	GB	United Kingdom	MC	Monaco	TD	Chad
BA	Bosnia and Herzegovina	GE	Georgia	MD	Republic of Moldova	TG	Togo
BB	Barbados	GH	Ghana	MG	Madagascar	TJ	Tajikistan
BE	Belgium	GN	Guinea	MK	The former Yugoslav Republic of Macedonia	TM	Turkmenistan
BF	Burkina Faso	GR	Greece			TR	Turkey
BG	Bulgaria	HU	Hungary	ML	Mali	TT	Trinidad and Tobago
BJ	Benin	IE	Ireland	MN	Mongolia	UA	Ukraine
BR	Brazil	IL	Israel	MR	Mauritania	UG	Uganda
BY	Belarus	IS	Iceland	MW	Malawi	US	United States of America
CA	Canada	IT	Italy	MX	Mexico	UZ	Uzbekistan
CF	Central African Republic	JP	Japan	NE	Niger	VN	Viet Nam
CG	Congo	KE	Kenya	NL	Netherlands	YU	Yugoslavia
CH	Switzerland	KG	Kyrgyzstan	NO	Norway	ZW	Zimbabwe
CI	Côte d'Ivoire	KP	Democratic People's Republic of Korea	NZ	New Zealand		
CM	Cameroon			PL	Poland		
CN	China	KR	Republic of Korea	PT	Portugal		
CU	Cuba	KZ	Kazakstan	RO	Romania		
CZ	Czech Republic	LC	Saint Lucia	RU	Russian Federation		
DE	Germany	LI	Liechtenstein	SD	Sudan		
DK	Denmark	LK	Sri Lanka	SE	Sweden		
EE	Estonia	LR	Liberia	SG	Singapore		

NETWORK ADAPTOR DRIVER WITH DESTINATION BASED ORDERING

BACKGROUND OF THE INVENTION

This application claims priority from provisional patent application serial number 60/032,124, filed December 5, 1996, which discussed a number of background concepts related to the invention.

The current invention relates to the field of electronic circuits. More particularly, the current invention relates most directly to improvements in networked computer environments and has particular applications to the transmission of information between digital devices over a communications medium. The invention also is concerned with the interface between a network adaptor and its host operating system in order to improve network performance and reduce network operation burden on a host processor. The invention concerns in some details a network adaptor driver, which generally consists of program code running on a host's CPU which controls or interfaces with aspects of adaptor operation.

The present invention has applications to the field of computer systems and networks. A very wide variety of types of computer systems and networks exist, each having variations in particular implementations. The present invention will be described with reference to particular types of systems for clarity but this should not be taken to limit the invention, and it will be apparent to those of skill in the art that the invention has applications in many different types of computer systems. The invention therefore should not be seen as limited except as specifically herein provided.

Digital computer networks have become ubiquitous in academic, industry, and office environments. A number of different aspects of computer networks are discussed in co-

assigned pending U.S. applications serial nos. 08/313,674; 08/542,157; 08/506,533; and 08/329,714 each of which are incorporated herein by reference.

Networking Devices Standards

This specification presumes familiarity with the general concepts, protocols, and devices currently used in LAN networking and WAN internetworking applications such as, for example, the IEEE 802 and ISO 8802 protocol suites and other series of documents released by the Internet Engineering Task Force that are publicly available and discussed in more detail in the above-referenced patent applications and will not be fully discussed here.

Fig. 1

Fig. 1 illustrates a local area network (LAN) 40 of a type that might be used today in a moderate-sized office or academic environment and as an example for discussion purposes of one type of network in which the present invention may be effectively employed. LANs are arrangements of various hardware and software elements that operate together to allow a number of digital devices to exchange data within the LAN and also may include internet connections to external wide area networks (WANs) such as WANs 82 and 84. Typical modern LANs such as 40 are comprised of one to many LAN intermediate systems (ISs) such as ISs 60-62 and 67 that are responsible for data transmission throughout the LAN and a number of end systems (ESs) such as ESs 50a-d, 51a-c, and 52a-g, that represent the end user equipment. The ESs may be familiar end-user data processing equipment such as personal computers, workstations, and printers and additionally may be digital devices such as digital telephones or real-time video displays. Different types of ESs can operate together on the same LAN. In one type of LAN, LAN ISs 60-61 are referred to as bridges and WAN ISs 64 and 66 are referred to as routers, and IS 67 is referred to as a repeater, however many different LAN configurations are possible, and the invention is not limited in application to the network shown in Fig. 1.

The LAN shown in Fig. 1 has segments 70a-e, 71a-e, and 72a-e, and 73a. A segment is generally a single interconnected medium, such as a length of contiguous wire, optical fiber, or coaxial cable or a particular frequency band. A segment may connect just two devices, such as segment 70a, or a segment such as 72d may connect a number of devices using a carrier sense multiple access/collision detect (CSMA/CD) protocol or other multiple access protocol such as a token bus or token ring. A signal transmitted on a single segment, such as 72d, is simultaneously heard by all of the ESs and ISs connected to that segment.

LANs also may contain a number of repeaters, which is one configuration possible shown for device 67. A repeater generally repeats out of each of its ports all data received on any one port, such that the network behavior perceived by ESs such as 50d-f is identical to the behavior they would perceive if they were wired on the same segment such as 52d-g. Repeaters configured in a star topology, such as 67, are also referred to as hubs. In alternative network topologies, device 67 could be a bridge as described below.

Drivers, Adaptors, and LAN Topology

Each of the ISs and ESs in Fig. 1 includes one or more adaptors and a set of drivers. An adaptor generally includes circuitry and connectors for communication over a segment and translates data from the digital form used by the computer circuitry in the IS or ES into a form that may be transmitted over the segment, e.g., electrical signals, optical signals, radio waves, etc. An ES such as 50b will generally have one adaptor for connecting to its single segment. A LAN IS such as 61 will have five adaptors, one for each segment to which it is connected. A driver is a set of instructions resident on a device that allows the device to accomplish various tasks as defined by different network protocols. Drivers are generally software programs stored on the ISs or ESs in a manner that allows the drivers to be modified without modifying the IS or ES hardware.

LANs may vary in the topology of the interconnections among devices. In the context of a communication network, the term "topology" refers to the way in which the stations attached to the network are interconnected. Common topologies for LANs are bus, tree, ring, and star. LANs may also have a hybrid topology made up of a mixture of these. The overall LAN pictured in Fig. 1 has essentially a tree topology, but incorporating one segment, 72d, having a bus topology, and incorporating one segment 70d having a star topology. A ring topology is not shown in Fig. 1, but it will be understood that the present invention may be used in conjunction with LANs having a ring topology.

Other Network Devices

The LAN ISs in LAN 40 include bridge/switches 60-63. Bridges are understood in the art to be a type of computer optimized for very fast data communication between two or more segments. A bridge according to the prior art generally makes no changes to the packets it receives on one segment before transmitting them on another segment. Bridges are not necessary for operation of a LAN and, in fact, in prior art systems bridges are generally invisible to the ESs to which they are connected and sometimes to other bridges and routers.

Packets

In a LAN such as 40, data is generally transmitted between ESs as independent packets, with each packet containing a header having at least a destination address specifying an ultimate destination and generally also having a source address and other transmission information such as transmission priority. ESs generally listen continuously to the destination addresses of all packets that are transmitted on their segments, but only fully receive a packet when its destination address matches the ES's address and when the ES is interested in receiving the information contained in that packet.

Fig. 2 depicts an example of a packet as it may be transmitted to or from router 64 on LAN segment 73a. The

example shown is essentially an Ethernet packet, having an Ethernet header 202 and a 48-bit Ethernet address (such as 00:85:8C:13:AA) 204, and an Ethernet trailer 230. Within the Ethernet packet 200 is contained, or encapsulated, an IP packet, represented by IP header 212, containing a 32 bit IP address 214 (such as 199.22.120.33). Packet 200 contains a data payload 220 which holds the data the user is interested in receiving or holds a control message used for configuring the network. Many other types and configurations of packets are known in the networking art and will be developed in the future.

Layers

An additional background concept important to understanding network communications is the concept of layered network protocols. Modern communication standards, such as the TCP/IP Suite and the IEEE 802 standards, organize the tasks necessary for data communication into layers. At different layers, data is viewed and organized differently, different protocols are followed, and different physical devices handle the data traffic. Fig. 3 illustrates one example of a layered network standard having a number of layers, which we will refer to herein as the Physical Layer, the Data Link Layer, the Routing Layer, the Transport Layer and the Application Layer. These layers correspond roughly to the layers as defined within the TCP/IP Suite. (The 802 standard has a different organizational structure for the layers and uses somewhat different names and numbering conventions.)

An important ideal in layered standards is the ideal of layer independence. A layered protocol suite specifies standard interfaces between layers such that, in theory, a device and protocol operating at one layer can coexist with any number of different protocols operating at higher or lower layers, so long as the standard interfaces between layers are followed.

Adaptor to Host Interface

Another aspect of networks is the interface between the network and the host operating system or processors that transmit data via the network. Some types of network protocols may require a large amount of attention from a host processor. This can be undesirable where a hosts activity on the network impinges on the hosts processor's ability to perform other host functions such as running user applications. Adaptors may also differ in their ability to buffer network traffic. Some adaptors rely on the host to buffer most network traffic and do not include a large amount of buffer memory on the adaptor itself.

Increasing Traffic Capacity of Some Network Devices Create a Need For New Solutions To Improve Network Performance

In recent years, the amount of data users wish to transmit over a network has increased dramatically. This increase has placed an increasingly heavy burdens on all parts of the network. A number of existing networks include a mixture of components or segments, some capable of operating at a maximal speed of the network and others operating at slower speeds.

What is needed is an improved network and components allowing for greater utilization of network resources in a distributed network environment and with techniques for preventing bottlenecks in slower components in a distributed network from unduly degrading overall network performance or preventing full utilization of faster components.

SUMMARY OF THE INVENTION

In general terms, the present invention comprises techniques and devices for an improved computer network and methods and devices for the operation thereof.

Reordering transmitted packets at a network adaptor

In this aspect, the present invention is a network adaptor driver with a destination based ordering scheme where

multiple packets of data are ordered before transmit to ensure fullest utilization of distributed network resources.

A network represents a distributed environment in which the various components (segments and devices) have differing ability to handle network traffic. The introduction of very high speed network components such as 100 Megabit Ethernet and Gigabit Ethernet increases the differences in ability of various parts of the network to handle data traffic. In addition, the ability of some network components to handle traffic varies over time. A network attached file server, for example, may communicate with large numbers of other network attached devices.

Modern network operating systems (OSs) such as Netware and Windows95 provide interfaces allowing protocols running within those OSs to request packet transmission from the underlying network hardware. In prior art systems, it is generally assumed that these packets will be transmitted by the network hardware in a FIFO order, with packets transmitted by the hardware in the order they are received from the protocols running on the OSs.

However, work in network bridges and switches has shown the FIFO ordering need only be preserved among packets from a particular source that are going to a particular destination (i.e. with a particular source/destination address pair) in order for the network to function as the protocols expect. For example, if node A is transmitting to nodes B and C then the order of packets from A to B must be preserved and the order of packets from A to C must be preserved, but it is not necessary to preserve the order of packet from A to B with respect to the packets from A to C.

Therefore it is possible to improve network parallelism at layer 2 and below by reordering packets within the adaptor hardware and driver, transparently to the higher layer protocols requesting transmission of those packets, in order to distribute packets to as many different ultimate destinations as possible. This avoids the transmitter waiting for slower network paths. This technique also can reduce bottlenecks developing further down the transmission stream in

network ISs. According to the invention, this redistribution is accomplished while still preserving the order of packets sent to any particular destination from any particular source. Thus, the invention effective spreads packets over a number of destinations at layer 2, without the need for modifying higher layer protocols.

Specific aspects of the invention will be better upon reference to the following description of specific embodiments and the attached claims.

BRIEF DESCRIPTION OF THE DRAWINGS

FIG. 1 is a diagram of a network of one type in which the invention may be effectively employed;

FIG. 2 is a diagram illustrating a server transmitting to three receivers via a network to illustrate aspects of the invention;

FIG. 3 is a diagram of a prior art packet as an example of a type of data unit upon which the invention may be effectively employed.

FIG. 4 is a diagram illustrating a layered network protocol.

FIG. 5 is a diagram of an adaptor on a network with indicated packet reordering according to two different embodiments of the invention.

DETAILED DESCRIPTION OF SPECIFIC EMBODIMENTS

According to the invention, an adaptor and/or adaptor driver reorder packets to be transmitted based on the destinations of those packets. The invention effectively distributes packets more evenly through various network transmission paths and thus increases overall network throughput. The invention also effectively reduces bottlenecks at the transmitter. Packets to be transmitted may be reordered according to the invention using any number of different destination address based ordering schemes, examples of which are discussed below.

Destination Address Based Ordering: Example 1

In one destination address based ordering scheme, all packets from the same source address destined for the same destination are placed in a queue created in system memory. Therefore, as packets are received from higher layer protocols and placed in memory for transmission, associated with each destination address is essentially a FIFO queue of packets addressed to that particular device. According to one approach, the invention samples all existing destination queues in memory, in round-robin fashion, and sends one (or a group) of packets from all queues containing packets, in the round-robin order. This approach distributes packets to all destination without regard to the order at which packets to different destinations were queued by the higher layer protocol.

As an example of this first approach, assume higher layer protocols submitted packets in following order for destinations A, B and C:

A1, A2, B1, C1, C2, C3, A3, B2 , B3

For "round-robin" destination based ordering the typical transmission order would be:

A1, B1, C1, A2, B2, C2, A3, B3, C3.

It will be seen that if, for example, there is a bottleneck in the path to C such that after one packet is transmitted to C the adaptor must wait a relative long time before transmitting another packet to C, the second ordering is more likely to get the most packets through the network in the least amount of time than the first ordering.

Destination Address Based Ordering: Example 2

In this embodiment, the invention attempts to preserve the original order of packets to varying destinations as much as possible while still distributing packets to all destinations. In other words, this is an attempt to reduce the variation between the time when packets are queued by a higher layer protocol and when the packets are actually transmitted on the network, which is a form of jitter. In this embodiment, a temporal window is defined in which a

packet from each destination will be selected for transmission. However, the order in which the packets within the temporal window are transmitted is determined by the temporal order that they were submitted to the adaptor driver or interface by the higher layer protocol.

Assume again a protocol or protocols submitted packets in following order for destinations A, B and C:

A1, A2, B1, C1, C2, C3, A3, B2, B3

For temporal window based ordering the typical transmission order would be:

A1, B1, C1 ¦ A2, C2, B2 ¦ C3, A3, B3

Observing the two transmission order examples, one can observe that the destination address based ordering creates a uniform packet ordering where the temporal window approach minimizes the time to transmission for a given packet while still distributing packets over all destinations.

Destination Address Based Ordering: Other Examples

Any number of other algorithms are possible for determining the exact transmission ordering of packets and different mechanisms may be chosen in different environments. Other embodiments are possible in which packets are delayed by no more than a certain amount of time before transmission on a network, despite the destination based ordering considerations. However, in network environments presently in contemplation, the one of the more simplified ordering schemes just described would be represent a preferred embodiment.

It is also possible that an adaptor driver according to the invention will include multiple alternative methods for determining the order for transmitting destination-spread packets, with a particular algorithm selectable by a user depending on the specific requirements of the network in which the adaptor is used.

The invention has now been explained with reference to specific and alternative embodiments. Other embodiments will be obvious to those of skill in the art. The invention therefore should not be limited except as provided for in the attached claims as extended by allowable equivalents.

WHAT IS CLAIMED IS:

1. A network adaptor driver comprising:

an interface for receiving data from a host;

an interface for transmitting packets of data over a network;

a mechanism for reordering packets of data received from said host based on a destination address of said packets before transmitting on said network.

2. The network adaptor driver according to claim 1 wherein said reordering is determined solely by the destination address of said packets, said reordering transmitting a number of packets over each queued destination address before transmitting the next number of packets over each said queued destination address.

3. The network adaptor driver according to claim 1 wherein said reordering is determined partly by the destination address of said packets and partly by when a packet is queued by said host so that packets are distributed over all destinations while minimizing the time to transmission from when a packet is received from the host for a given packet.

4. The network adaptor driver according to claim 1 wherein, for example, packets received from the host in an order A1, A2, B1, C1, C2, C3, A3, B2, B3 are transmitted by the adaptor in nearly the order A1, B1, C1, A2, B2, C2, A3, B3, C3.

5. The network adaptor driver according to claim 1 wherein, for example, packets received from the host in an order A1, A2, B1, C1, C2, C3, A3, B2, B3 are transmitted by the adaptor in nearly the order A1, B1, C1, A2, C2, B2, C3, A3, B3.

6. A method for maximizing network parallelism comprising:

receiving data packets for transmit to a plurality of destinations from a host in a first FIFO order;

prior to transmitting said data packets, reordering said packets based on a destination address of said packets, so that said packets are spread over different network destination paths; and

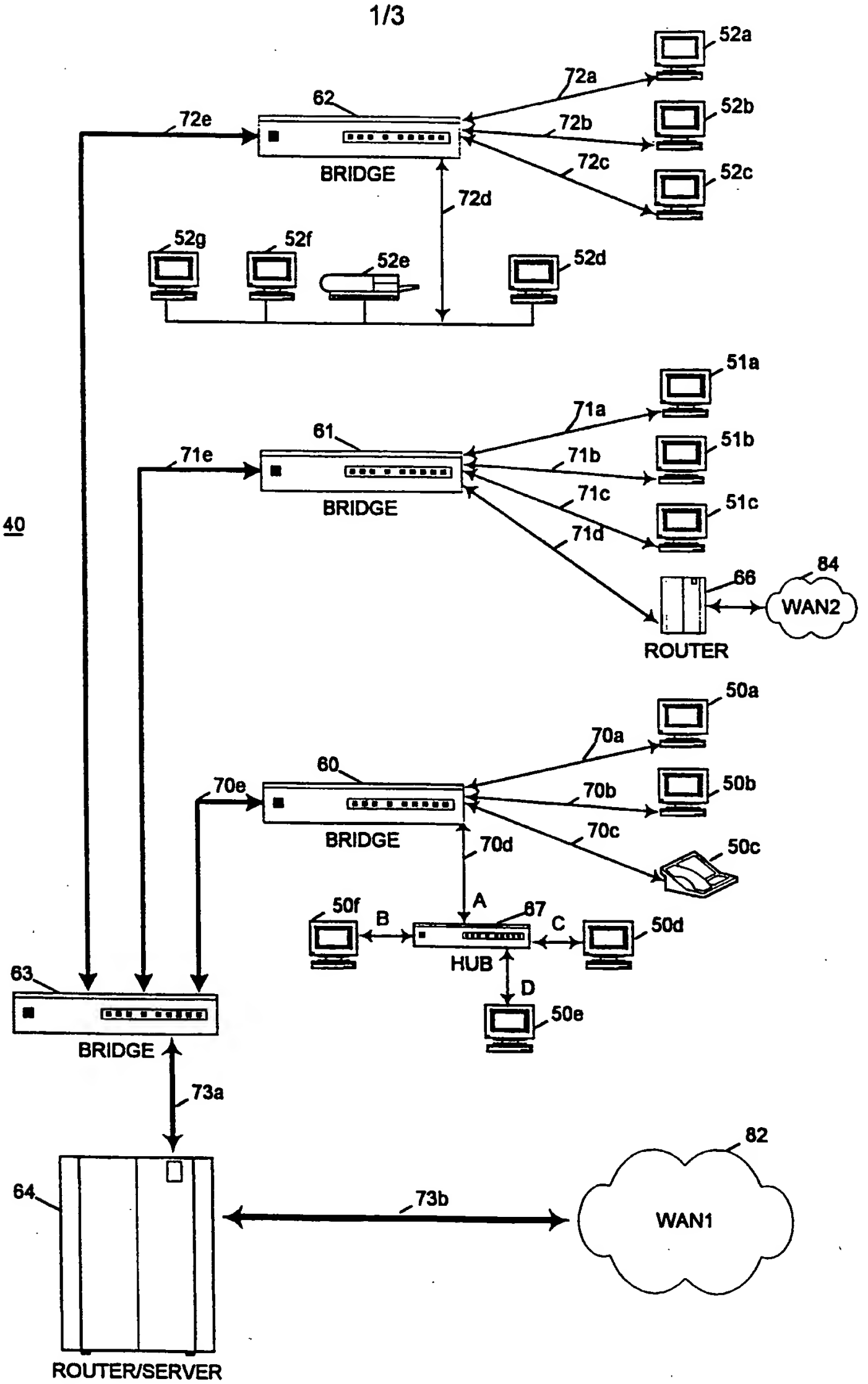
transmitting said reordered packets over a network.

7. The method according to claim 6 wherein said reordering is determined solely by the destination address of said packets without regard to the time at which packets to different destinations are queued.

8. The method according to claim 7 wherein said reordering is determined partly by the destination address of said packets and partly by when a packet is received from said host so that packets are distributed over all destinations while minimizing the time variations between when a packet is received from the host for a given destination and when that packet is transmitted.

9. The network adaptor driver according to claim 1 wherein said reordering is determined by a preset, nonadjustable scheme.

10. The network adaptor driver according to claim 1 wherein said reordering is determined by a programmable scheme which may take into account differences in speed and performance paths to particular destinations to maximize network parallelism.


52a
72a
62
72e
52b
72b
BRIDGE
72c
52c
72d
52g
52f
52e
52d
51a
71a
61
71e
51b
71b
BRIDGE
71c
51c
71d
66
84
WAN2
ROUTER
50a
70a
60
70e
50b
70b
BRIDGE
70d
70c
50c
A
50f
B
87
C
50d
HUB
63
D
50e
BRIDGE
73a
82
64
73b
WAN1
ROUTER/SERVER

40

FIG. 1

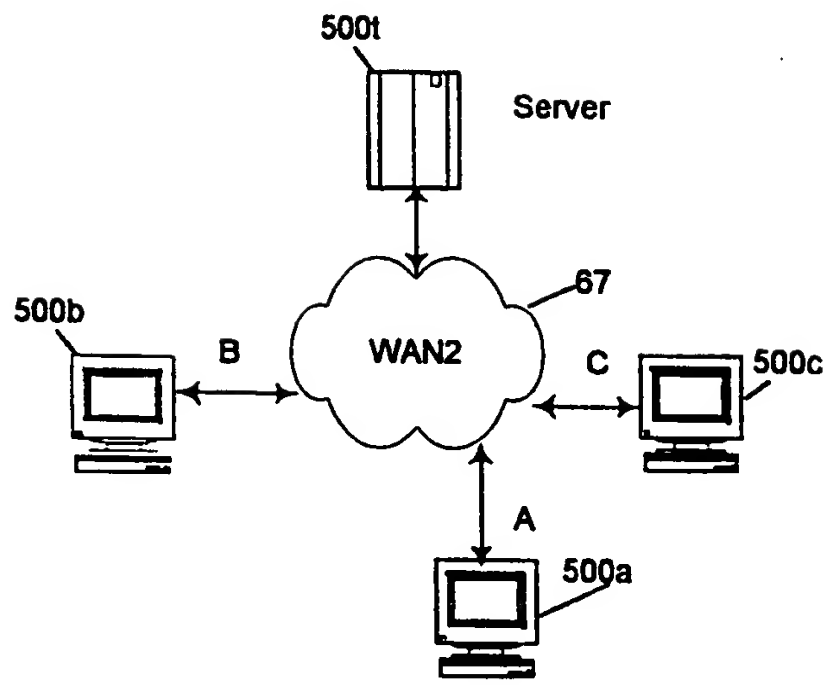


FIG. 2

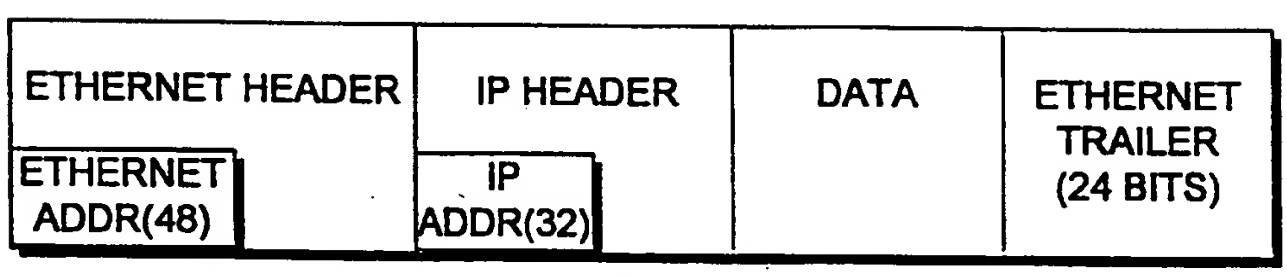


FIG. 3A

HIGH

LAYER NAME (NUMBER)	DEVICES	DATA	PROTOCOLS
HIGHER LAYER PROTOCOLS			
APPLICATION LAYER (5)		FILES	FTP, HTTP
TRANSPORT LAYER (4)	ROUTERS	ROUTING PACKETS	TCP, UDP
ROUTING LAYER (3)	ROUTERS	ROUTING PACKETS	IP
DATA LINK LAYER (2)	BRIDGES	PACKETS	ETHERNET
PHYSICAL LAYER (0,1)	REPEATERS	BITS	ETHERNET

LOW

FIG. 4

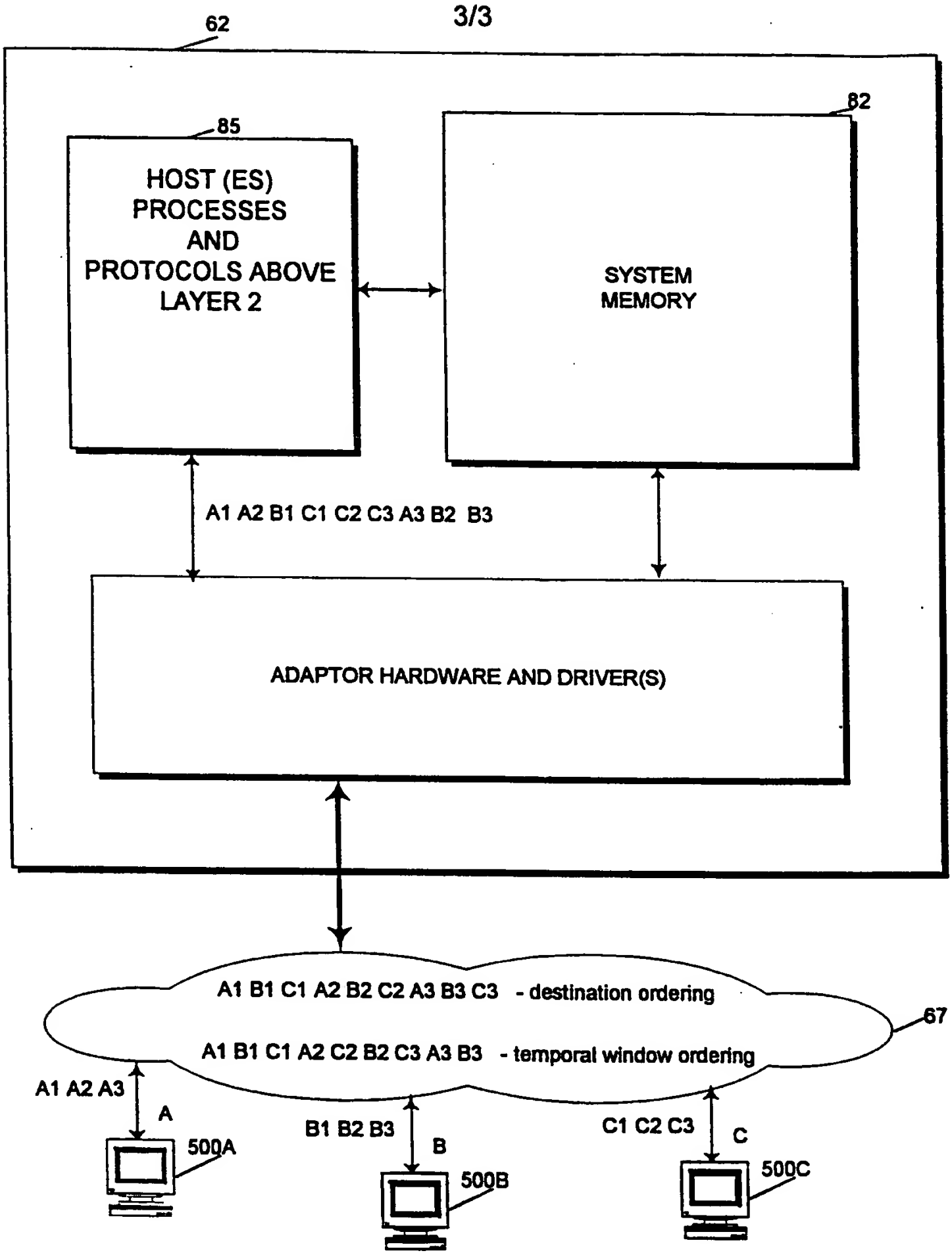

62
82
85
HOST (ES) PROCESSES AND PROTOCOLS ABOVE LAYER 2
SYSTEM MEMORY
A1 A2 B1 C1 C2 C3 A3 B2 B3
ADAPTOR HARDWARE AND DRIVER(S)
A1 B1 C1 A2 B2 C2 A3 B3 C3 - destination ordering
A1 B1 C1 A2 C2 B2 C3 A3 B3 - temporal window ordering
67
A1 A2 A3
A
500A
B1 B2 B3
B
500B
C1 C2 C3
C
500C

FIG. 5

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US97/22620

A. CLASSIFICATION OF SUBJECT MATTER

IPC(6) :H04L 12/56
US CL :370/392, 393, 394, 412, 415, 416, 428, 429

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

U.S. : 370/392, 393, 394, 412, 415, 416, 428, 429

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

APS, IEEEONDISC
search terms: round robin, fair, queue, destination, fifo

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	MORGAN, S.P. "Queueing Disciplines and Passive Congestion Control in Byte-Stream Networks" IEEE INFOCOM'89 The Conference on Computer Communications. April 1989. Vol. 2, pages 711-720, especially pages 711, 713 and 714.	1-10
Y	CHERKASOVA, L. ET AL. "The Impact of Message Scheduling on a Packet Switching Interconnect Fabric", Proceedings of the Twenty-Ninth Hawaii International Conference on System Sciences". January 1996. Vol. 1, pages 668-669.	1-10

☐ Further documents are listed in the continuation of Box C. ☐ See patent family annex.

*	Special categories of cited documents:	"T"	later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"A"	document defining the general state of the art which is not considered to be of particular relevance	"X"	document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"E"	earlier document published on or after the international filing date	"Y"	document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
"L"	document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)	"&"	document member of the same patent family
"O"	document referring to an oral disclosure, use, exhibition or other means		
"P"	document published prior to the international filing date but later than the priority date claimed		

Date of the actual completion of the international search	Date of mailing of the international search report
24 MARCH 1998	06.05.98
Name and mailing address of the ISA/US Commissioner of Patents and Trademarks Box PCT Washington, D.C. 20231 Facsimile No. (703) 305-3230	Authorized officer MELVIN MARCELO Telephone No. (703) 305-3900

Form PCT/ISA/210 (second sheet)(July 1992)★